Journal of Applied Statistics, Vol. 29, Nos. 1-4, 2002, 241-244

Response to Carl Schwarz

J. A. DUPUIS, *Laboratoire de Statistique et Probabilités, University Paul Sabatier, France*

I would like to thank Carl Schwarz for his thoughtful comments as well as for his stimulating questions.

Question 1

We think that the state-space formulation (suggested by Schwarz) is not appropriate to describe the missing data phenomenon inherent in multi-strata capture-recapture data. Contrary to the state-space models the observation process (that corresponds to the capture process in our paper) has no observation error. When animal *i* has been captured at time *t*, its position (and thus the observation) is known without error; but there is no observation error when animal *i* has not been captured at time *t*. Its position is simply not available from the data, and it is simply missing. Finally, we think that the Arnason-Schwarz model is typically a missing data model, in the same way as the mixture models or the hidden Markov chains models.

Question 2

As pointed out by Schwarz, problems of non-identifiability are crucial for frequentist statisticians. From a Bayesian point of view, non-identifiable parameters can be estimated, provided the posterior distribution exists. This interesting characteristic has been stressed by Brooks *et al.* (2000) when analysing, from a Bayesian point of view, a ring-recovery data set, for which the likelihood has a completely flat ridge. In that situation, there is no unique MLE, whereas the Bayesian estimate exists and is remarkably precise.

Correspondence: Laboratoire de Statistique et Probabilités, University Paul Sabatier, Bat. 1R1, 118 Route de Narbonne, 31062 Toulouse, France. E-mail: dupuis@cict.fr

ISSN 0266-4763 print; 1360-0532 online/02/010241-04

DOI: 10.1080/02664760120108656

TABLE 1

	ϕ_1	ϕ_2	p_2	p_3
mean	0.53	0.67	0.39	0.67
SD	0.09	0.17	0.07	0.17

TABLE 2

	$\psi(1, 1)$	$\psi(2, 2)$	$p(1)$	$p(2)$
mean	0.598	0.786	0.500	0.749
SD	0.035	0.064	0.039	0.042

Schwarz suggests that Bayes estimates could be used, firstly, to understand how information contained in the data is parcelled between two (or more) non-identifiable parameters and, secondly, to diagnose which parameters are identifiable and which parameters are not. To my knowledge, no work has been devoted to investigating such issues. To examine those questions we consider two models for which some parameters are not identifiable: the first one is the CJS (Cormack-Jolly-Seber) model, and the second one is the AS$_h$ (time-homogeneous closed Arnason-Schwarz) model. For each of these two models we calculate the Bayesian estimates of non-identifiable parameters (uniform prior having been put on all the parameters).

We first consider a data set that is constructed as in Section 6 of our paper. The experimental protocol includes $T = 3$ capture-recapture sessions (including the tagging session); tagging has been carried out only at time $t = 1$. The model is parameterized by $\theta = (\phi_1, p_2, \phi_2, p_3)$. Expected counts are calculated under the CJS model, and for a fixed value $\theta^\star$ of θ. For $\theta^\star = (0.5, 0.4, 0.75, 0.6)$ and $n = 200$, it is easy to check that the counts are for each history: 111(**18**), 101(**27**), 110(**22**), 100(**133**). It is well known that ϕ_2 and p_3 are not identifiable whereas the product $\beta_3 = \phi_2 p_3$ is identifiable. We denote by $\hat{\theta}$ the Bayesian estimate of θ. Table 1 provides the posterior means and the posterior standard deviations (SDs) of the parameters.

We now construct a data set under the AS$_h$ model. The study zone K includes two strata 1 and 2, and the experimental protocol includes $T = 3$ capture-recapture sessions (including the tagging session); tagging is carried out only at time $t = 1$, and only in stratum 1. The AS$_h$ model is parameterized by $\theta = (\psi(1, 1), \psi(2, 2), p(1), p(2))$. Expected counts are calculated under the AS model, and for a fixed value $\theta^\star$ of θ. For $\theta^\star = (0.6, 0.8, 0.5, 0.75)$ and $n = 200$, it is easy to check that the counts are for each history: 122(**36**), 121(**6**), 120(**18**), 111(**18**), 112(**18**), 110(**24**), 101(**20**), 102(**30**), 100(**30**). Table 2 provides the posterior means and the posterior standard deviations (SDs) of the parameters.

We now examine the identifiability of the AS$_h$ model. Let $\theta = (\psi(1, 1), \psi(2, 2), p(1), p(2))$ and $\theta' = (\psi'(1, 1), \psi'(2, 2), p'(1), p'(2))$ so that:

$$\psi(1, 1)p(1) = \psi'(1, 1)p'(1) \qquad \psi(2, 2)p(2) = \psi'(2, 2)p'(2)$$

and $\psi(1, 1) + \psi(2, 2) = \psi'(1, 1) + \psi'(2, 2) = 1$. It is easy to check (Dupuis, 2001), that for all data sets $\mathbf{y}$, we have $L(\theta|\mathbf{y}) = \mathbf{L}(\theta'|\mathbf{y})$. Therefore the AS$_h$ model is not identifiable. Note that $\psi(1, 1) + \psi(2, 2) \neq 1$ is satisfied if and only if the movement process is Markovian, in the strict sense (that is, if we have $\psi(1, 1) \neq \psi(2, 1)$).

In the first example (associated with the CJS model), Bayesian estimates of the identifiable parameters (i.e. of ϕ_1 and p_2) are satisfactory. By comparison, $\hat{\phi}_2$ and $\hat{p}_3$ are clearly erroneous. Note that $\hat{\phi}_2 = \hat{p}_3$. This is not surprising, since ϕ_2 and p_3 appear in the likelihood only through the product $\phi_2 p_3$. Note also that $\hat{\phi}_2 \hat{p}_3 = 0.67^2 \simeq 0.45$ is very close to $\beta_3^\star = 0.45$. The Bayesian estimates of ϕ_2 and p_3 are therefore illusory since they do not provide information on the value of each parameter separately. The only information provided by $\hat{\phi}_2$ and $\hat{p}_3$ is related to the product $\phi_2 p_3$. Nevertheless, if some prior information is available on ϕ_1, p_2 and p_3, the parameterization $\theta = (\phi_1, p_2, \phi_1, p_3)$ can be relevant. In our example, if $\phi_1 \sim \mathscr{B}e(10, 10)$; $p_2 \sim \mathscr{B}e(8, 12)$; and $p_3 \sim \mathscr{B}e(12, 8)$, we obtain $\hat{\phi}_2 = 0.74$ with a posterior SD equal to 0.12.

In the second example, all the Bayesian estimates are precise, unlike the first example. This is not surprising, since θ is identifiable if and only if $\psi(1, 1) \neq \psi(2, 1)$, and we have chosen $\theta^\star$ such as $\psi^\star(1, 1) \neq \psi^\star(2, 1)$. When the data set is constructed under the AS$_h$ model with $\psi^\star(1, 1) = \psi^\star(2, 1)$, the Bayesian estimate of θ is completely erroneous. For example, for $\theta^\star = (0.2, 0.8, 0.5, 0.75)$ and $n = 500$, we have obtained $\hat{\theta} = (0.40, 0.60, 0.26, 0.99)$ and 0.02, 0.028, 0.022 and 0.0010, for the respective posterior SDs. From those results, we can observe a phenomenon similar to the one observed with the CJS model: $\widehat{\psi(1, 1)}\widehat{p}(1)$ is very close to $\psi(1, 1)p(1)$ (idem for $\widehat{\psi(2, 2)}\widehat{p}(2)$ and $\psi(2, 2)p(2)$).

Results associated with these two examples (as well as some additional analyses not reported here) seem to indicate that Bayesian estimates of non-identifiable parameters can have abnormally large posterior SD (but not necessarily, see the second example), and are generally very sensitive to the prior. Moreover, we have observed that the presence of non-identifiable parameters can reduce the speed of convergence of the Gibbs sampling. Results associated with these two examples also suggest the following procedure to detect non-identifiable parameters in capture-recapture models. Under the considered model assumed to be parameterized by $\theta = (\theta_j; j = 1, \ldots, \mathcal{J})$, construct a data set based upon the expected counts from a large sample (expected counts being calculated for a fixed value $\theta^\star$ of θ). The abnormal distance between $\theta_j^\star$ and its non-informative Bayesian estimate $\hat{\theta}_j$ should indicate that θ_j is not identifiable.

Question 3

The question of Schwarz leads us to provide some additional information concerning the construction of a prior distribution on a multi-dimensional parameter $\theta = (\theta_j; j = 1, \ldots, \mathcal{J}\}$ where $\mathcal{J} > 3$, $\theta_j \in]0, 1[$ and $\Sigma_{j=1}^{\mathcal{J}} \theta_j = 1$. This context concerns the incorporation of some prior information on movement parameters when the study zone has been divided in three zones (or more). In our paper, we have advocated the use of a Dirichlet distribution when the prior information consists of a prior mean μ_j for each component θ_j and of a 95% prior credible interval I for one of the components of the vector θ. When the prior information consists of a prior mean μ_j for each component θ_j and of a prior credible interval I_j for each component θ_j, the Dirichlet distribution is clearly not able to incorporate such a prior. In that case, we advocate the following conservative attitude (suggested by Schwarz). Search λ_j so that $\Pr(\theta_j \in I_j) = 0.95$ assuming that $\theta_j \sim \mathscr{B}e(\lambda_j \mu_j, \lambda_j(1 - \mu_j)$. Then take $\lambda = \inf_j \lambda_j$, accepting a certain loss of precision of the prior information. Finally $\theta \sim \mathscr{D}(a_1, \ldots, a_j, \ldots, a_j)$ where $a_j = \lambda \mu_j$.

Question 4

The data set is the one considered in Section 6.1. Moreover, as in Section 6, we only focus on the parameter $\psi_1(2,1)$. Unless otherwise specified, we have put a uniform prior distribution on all the parameters.

The MLE of $\psi_1(2,1)$ given by MARK (that is, 0.56), is very close to the non-informative Bayesian estimation (that is, 0.57). When we compare the frequentist *CI* yielded by MARK and the Bayesian *CI*, the comparison clearly shows the advantage of the Bayesian *CI* (see Section 6.1). It is not due to some additional information incorporated via the prior (as suggested by Schwarz), since a uniform prior distribution has been put on $\psi_1(2,1)$. Actually, I think that the frequentist *CI* yielded by MARK is so wide because its construction relies on assumptions of asymptotic normality (which requires the size n of the sample to be large), whereas in our example n is relatively small (since $n = 40$). The Bayesian *CI* we provide is exact.

Schwarz suggests that the formula (4), that is

$$\hat{\theta}_\pi = \frac{n}{\lambda + n}\hat{\theta}_{ml} + \frac{\lambda}{\lambda + n}\mathbb{E}[\theta]$$

could be used to indicate the (relative) amount of information in the posterior from each source. Unfortunately, in the AS model, this formula does not apply to $\psi_t(r,s)$, except if the locations between times t and $t+1$ have been all observed. If, for example $\psi_1(2,1) \sim \mathscr{B}e(0.7, 10)$, we have $\mathbb{E}[\psi_1(2,1)|\mathbf{y}] = 0.66$ (see Table 1 in our paper), whereas $0.8 \times 0.56 + 0.2 \times 0.7 = 0.588$. Nevertheless, it is possible to search for the weights p_1 and $p_2 = 1 - p_1$, such as $\hat{\theta}_\pi = p_1\hat{\theta}_{ml} + p_2\mathbb{E}[\theta]$, where, for convenience, $\psi_1(2,1)$ has been denoted by θ, which yields $p_1 \simeq 0.3$ and $p_2 \simeq 0.7$. The weight p_1 can be interpreted as the (relative) amount of information provided by the data, and the weight p_2 as the (relative) amount of information provided by the prior information. It is of interest to compare $p_1 = 0.3$ with $n/(\lambda + n) = 0.8$. Note that $n/(\lambda + n)$ can be interpreted as the relative amount of information about the parameter $\psi_1(2,1)$, provided by a (virtual) data set of size $n = 40$ in which all the transitions starting from 1 at time $t = 1$ would have been all observed.

REFERENCES

Brooks, S. P., Catchpole, E. A., Morgan, B. J. T. & Barry, S. C. (2000) On the Bayesian analysis of ring-recovery data, *Biometrics*, 56, pp. 951-956.

Dupuis, J. A. (2001) Maximum likelihood estimate of transition probabilities from capture-recapture data (submitted).